Calculation of sample size for stroke trials assessing functional outcome: comparison of binary and ordinal approaches

The Optimising Analysis of Stroke Trials (OAST) Collaboration

Correspondence to:

Professor Philip Bath

Stroke Trials Unit

University of Nottingham

Clinical Sciences Building

City Hospital campus

Nottingham NG5 1PB UK

Tel: +44 115 823 1765

Fax: +44 115 823 1767

E-mail: philip.bath@nottingham.ac.uk

ACKNOWLEDGEMENTS

Writing committee: Laura J Gray (lead statistician, Nottingham, UK); Philip MW Bath (chief investigator, Nottingham, UK); Timothy Collier (statistical advisor, London, UK). We declare no conflicts of interest.

The following contributors provided individual patient data from their trial, and commented on the draft manuscript:

Abciximab: H Adams (USA), E Barnathan (USA); W Hacke (Germany),

ASK: G Donnan (Australia)

ASSIST 07 & 10: S Davis (Australia)

ATLANTIS A & B: G Albers, S Hamilton (USA)

BEST Pilot & Main: D Barer (UK)

Citicoline 1, 7, 10, 18: A Davalos (Spain)

Corr: S Corr (UK)

Dover Stroke Unit: P Langhorne (UK)

DCLHb: P Koudstaal, R Saxena (Netherlands)

Ebselen: T Yamaguchi (Japan)

ECASS II: W Hacke, E Bluhmki (Germany)

Factor VII: S Mayer (USA), K Begtrup (Denmark)

FISS: R Kay (Hong Kong)

FOOD 3: M Dennis (UK)

Gilbertson: L Gilbertson (UK)

INWEST: N-G Wahlgren, N Ahmed (Sweden)

IST: P Sandercock (UK), S Lewis (UK)

Kuopio Stroke Unit: J Sivenius (Finland)

Logan: P Logan (UK)

MAST-I: L Candelise (Italy), J Wardlaw (UK)

Newcastle Stroke Unit: H Rodgers (UK)

NINDS: J Marler (USA)

Parker: C Parker (UK)

Nottingham Stroke Unit: N Lincoln, P Berman (UK)

RANNTAS I & II, STIPAS, TESS I & II: P Bath (UK), B Musch (USA)

Walker 1 & 2: M Walker (UK)

Young: J Young, A Forster (UK)

We thank the patients who took part in these studies, and the trialists who shared their data. The study was conceived, initiated, managed, analysed, and interpreted independently of any pharmaceutical company. Each collaborator listed above commented on the draft manuscript.

FUNDING

LJG is funded, in part, by Medical Research Council, BUPA Foundation and The Stroke Association (UK). PB is Stroke Association Professor of Stroke Medicine. The funding sources had no involvement in this project.

TITLE PAGE

Calculation of sample size for stroke trials assessing functional outcome: comparison of binary and ordinal approaches

Cover title: Calculation of sample size for stroke trials

Tables:

TABLE 1. Comparison of sample sizes produced by 5 methods.

TABLE 2. Comparison of sample sizes using 4 methods of calculation relative to the proportion method for a good outcome (modified Rankin Scale $\leq$2 or Barthel Index $\geq$60) with results subcategorised by intervention.

Figures:

FIGURE 1a. Sample size comparisons at varying levels (β) of power for the IST trial of aspirin.

FIGURE 1b. Sample size comparisons at varying levels (β) of power for a trial of edaravone.

FIGURE 1c. Sample size comparisons at varying levels (β) of power for a trial of intra-arterial prourokinase.

Keywords: stroke; randomised controlled trial; statistical analysis; sample size; power

ABSTRACT

Introduction: Many acute stroke trials have given neutral results. Sub-optimal statistical analyses may be failing to detect efficacy. Methods which take account of the ordinal nature of functional outcome data are more efficient. We compare sample size calculations for dichotomous and ordinal outcomes for use in stroke trials.

Methods: Data from stroke trials studying the effects of interventions known to positively or negatively alter functional outcome - Rankin Scale and Barthel Index - were assessed. Sample size was calculated using comparisons of proportions, means, medians (according to Payne), and ordinal data (according to Whitehead). The sample sizes gained from each method were compared using Friedman 2 way ANOVA.

Results: 55 comparisons (54,173 patients) of active versus control treatment were assessed. Estimated sample sizes differed significantly depending on the method of calculation ($p<0.0001$). The ordering of the methods showed that the ordinal method of Whitehead and comparison of means produced significantly lower sample sizes than the other methods. The ordinal data method on average reduced sample size by 28% (inter-quartile range 14% to 53%) compared to the comparison of proportions; however, a 22% increase in sample size was seen with the ordinal method for trials assessing thrombolysis. The comparison of medians method of Payne gave the largest sample sizes.

Conclusions: Choosing an ordinal rather than binary method of analysis allows most trials to be, on average, smaller by approximately 28% for a given statistical power. Smaller trial sample sizes may help by reducing time to completion, complexity, and financial expense. However, ordinal methods may not be optimal for interventions

which both improve functional outcome and cause hazard in a subset of patients, e.g. thrombolysis.

INTRODUCTION

The majority of stroke trials assessing efficacy have reported neutral results.[1] There are many possible reasons for this including the use of suboptimal methods for analysing the primary outcome.[2] Most stroke trials use a measure of dependency as the primary outcome, this being assessed with a functional scale such as the Barthel Index (BI) or modified Rankin Scale (mRS). Scales such as the BI and mRS are ordinal in nature, for example the mRS has seven levels ranging from 0 (no symptoms at all) to 6 (death);[3] these categories have a natural ordering although the difference between the categories is not linear, i.e. the difference between a score of 3 (slight disability) and 4 (moderately severe disability) does not have the same magnitude as the difference between 0 (no symptoms at all) and 1 (no significant disability).[3] Historically, many trials have combined these ordered categories into two groups to create a binary end point i.e. comparing independence with combined death and dependence. Combining data in this way generally loses statistical power since data not crossing the binary cut point are effectively discarded. We have shown that statistical tests that use the original ordered categories describing dependency are statistically more efficient than those which dichotomise the data;[2] suitable approaches include ordinal logistic regression, the t-test, and the robust rank test (a variant of the Mann-Whitney U test). Importantly, the use of tests which analyse ordered categorical data do not assume linearity in the mRS, or a particular range of baseline stroke severity.

If the analysis of stroke trials should be changed from using dichotomous to polytomous functional outcome data, then it is important to consider how sample size should be calculated. Sample size estimation is an important part of trial design and is now a compulsory element when applying for funding and publishing completed

trials.[4, 5] Key components in any sample size calculation include the intended power (1-ß) and significance (α), and expected treatment effect.[6]

This paper compares sample size estimations obtained using different methods based on dichotomous and ordinal outcomes and using data from the 'Optimising the Analysis of Stroke Trials' (OAST) project.[2]

METHODS

'Optimising the Analysis of Stroke Trials' data

A detailed description of the OAST data set has been published.[2] In summary, we sought individual patient data from randomised controlled trials assessing functional outcome after stroke for interventions which were either positive or negative according to the trial publication, or were included in a meta analysis which showed overall benefit or harm; neutral trials in a neutral meta-analysis were excluded. Demographic (age, gender), trial (setting, intervention, length of follow up, result), patient severity, and functional outcome (BI,[7] mRS,[3] '3 question' scale [3Q, a 4-level derivative of the 7-level mRS] [8]) data were collected for each trial. In factorial trials or those having more than two treatment groups, data were analysed for each comparison of active therapy versus control. Where outcome data were scored at several time points (e.g. 1, 3 and 6 months) the time point used for the primary outcome was included. Data were shared by investigators or extracted from publications. Interventions included thrombolysis, anticoagulation, antihypertensives, antiplatelets, feeding, neuroprotection, occupational therapy, procoagulants and stroke units.

Sample size estimation

Four methods of sample size estimation were chosen for comparison; one is based on the proportion of events and is currently used in many acute stroke trials. The other three estimate sample size for ordinal or continuous outcomes.[2] All the methods of sample size estimation assume that the treatment groups are of equal size. In all cases z_α and z_β are the appropriate values from the standard Normal distribution based on the significance level (α) and power ($1-\beta$) chosen by the investigator. The methods of sample size estimation used are described in more detail in Appendix 1. None of the methods take into account drop out or non compliance and it is

customary to inflate any given sample size by around 10% to take account of these factors.

Comparison of methods

Each method of sample size estimation was carried out on each data set. The parameters needed within the calculation of each sample size were derived from each data set and then these were used to calculate the sample size needed as if these treatment effects were desired. The comparison of proportions method was carried out twice using two different definitions of a functional outcome: (i) death or poor outcome (BI <60, mRS 3-6, 3Q 1/2) vs. good outcome (BI 60-100, mRS 0-2, 3Q 3/4); (ii) death or poor outcome (BI <95, mRS 2-6, 3Q 1-3) vs. excellent outcome (BI 95/100, mRS 0/1, 3Q 4), see [2] for definitions of outcomes for the other scales used. This reflects that most trials historically used the poor/good outcome whilst recently there has been a tendency to rely on the poor/excellent outcome (largely based on the results of the NINDS tPA trial [9]).

In all cases significance was set at 5% with a power of 90%. The use of a fixed power of 90% will have ensured that the risk of a false negative was held constant. These sample sizes were then ordered within each trial and given a rank, with the lowest rank given to the method which produced the smallest sample size. A two-way analysis of variance test was then used to see on average which method had produced the lowest ranks and therefore the lowest sample sizes. We were then able to order the methods in terms of the average sample sizes given using Duncan's multiple range test. [10] Each method of sample size calculation was then compared to the proportion method for a 'good outcome' (as this is the current standard method used in stroke trials). The median multiplier by type of intervention was then

calculated, i.e. a value <1 shows that the method produces a smaller sample size than the proportion method and >1 shows that a larger sample size will result. Analyses were carried out in SAS (version 8.2), Stata (version 7) and GenStat (version 8.1, for the methods of Payne and Whitehead[11, 12]) and significance was taken at $p<0.05$.

RESULTS

Trials characteristics

The characteristics of the OAST data set have been published.[2] A total of 55 comparisons of active versus control treatment (54,173 patients) were included, these comprising individual patient data from 38 trials and summary data extracted from the publications of a further 9 studies; six trials had two active treatment groups, and one had three active groups so a further 8 comparisons were available. The data related to 34 acute stroke trials, 7 trials of rehabilitation (1,164 patients) and 6 trials of stroke units (1,399 patients). BI was used to measure functional outcome in 22 trials,[7] 18 used the mRS,[3] 3 used the 3Q scale,[8] 1 used the Rivermead scale, 2 related trials used the Nottingham ADL scale, and 1 trial used its own ordinal measure.

Comparison of sample size methods

The sample size methods differed significantly in the estimated sample sizes they produced for each trial ($p<0.0001$). The ordering of the methods showed that the ordinal method of Whitehead [11] and comparison of means method produced significantly lower sample sizes than the other approaches, with the comparison of medians method of Payne[12] giving the largest sample sizes (table 1). Table 2 shows the change in sample size in relation to the current standard method based on comparison of proportions for a good outcome (mRS $\leq$2 or BI $\geq$60). The ordinal method of Whitehead [11] and comparison of means appear to reduce sample size by 28% and 30% respectively relative to comparison of proportions (table 2). In contrast, the method of Payne[12] produces 12% larger sample sizes. Whilst this finding appears to be true for most interventions, it may not be correct for trials of thrombolysis where ordinal (Whitehead, Payne [11, 12]) and continuous (comparison of means) approaches produce larger sample sizes, interestingly, comparison of

proportions based on an 'excellent' outcome also led to an increase in sample size as compared with comparisons based on a 'good' outcome.

Figure 1 gives examples of the sample size required with varying levels of statistical power for each method for three trials with published summary data.[13-15] In the first two examples (aspirin, edaravone [13, 14]), the sample size produced according to Whitehead [11] gave smaller trials irrespective of power. In contrast, ordinal or continuous methods gave larger trials than for use of a binary outcome for the thrombolytic agent.[15]

DISCUSSION

The results support the contention that trials designed to use an ordinal analysis of functional outcome [2] will, on average, be smaller than those using a dichotomous outcome. In particular, Whitehead's method,[11] which assumes trials will be analysed using ordinal logistic regression, produces sample sizes which are typically 28% smaller than the dichotomous approach based on comparison of good outcome (mRS $\leq$2 or BI $\geq$60) (table 2, figures 1a and 1b). A similar reduction is seen using the comparison of means. Taking this finding with the results of the first OAST project,[2] we suggest, with one exception (see below), that stroke trialists should consider designing and analysing most trials using approaches which maintain the ordered categorical nature of functional outcome data based on mRS and BI. Analysis of means may be appropriate for polytomous outcomes with 7 or more levels,[16, 17] as occurs with the BI.

Ordinal logistic regression assumes the intervention will exert effects of similar magnitude and direction at each transition of the outcome scale, i.e. 'proportionality of odds'. This is unlikely to be the case for treatments where symmetrical benefits occur (i.e. the intervention is effective across a spectrum of severity) but hazard is asymmetrical tending to effect mainly those with severe stroke. Thrombolysis is an example and its overall effect is to reduce dependency and, to a lesser extent, increase death (largely through promoting fatal intracerebral haemorrhage).[18] Specifically, thrombolysis probably reduces dependency across all levels of the mRS, but increases haemorrhage in patients with severe stroke who are likely to have a poor outcome. Hence, thrombolysis may be considered, in the context of stroke severity, to have symmetrical effects on efficacy but asymmetrical effects on hazard. This is evident in table 2 and figure 1c where the ordinal (Whitehead, Payne[11, 12]) and

continuous methods did not deliver smaller thrombolysis trials, e.g. PROACT II.[15] In contrast, most other interventions are likely to move patients up (efficacy) or down (hazard) by a part (or whole) of a mRS level [2] therefore fulfilling the key assumption underlying proportionality of odds; table 2 shows that the ordinal method of Whitehead [11] leads to smaller sample sizes for a wide range of interventions including antiplatelets, neuroprotectants, occupational therapy, and stroke units. By example, the data for the pilot factor VIIa (FAST [19]) had symmetrical effects on both benefit (reduction in haematoma volume) and hazard (increase in ischaemic stroke and myocardial infarction) so that ordinal approaches appeared to be superior to those which dichotomise functional outcome.

The advantage of our study is that the different methods for estimating sample size have been tested on data from a large number of real stroke trials. As a result, the findings are likely to exhibit external validity. It is evident that stroke trials are inherently heterogeneous in their design and results; interventions, patients and results differ. Modelling approaches which synthesise data or use data from a single study cannot adequately take account of this heterogeneity. However, we were unable to obtain data for all the trials which fulfilled the study's inclusion criteria (see [2] for a list) thereby weakening the precision of our findings. A disadvantage of this study is that we aimed to include data from all stroke trials assessing a beneficial or harmful intervention. Unfortunately, data were not made available for all identified trials; where possible, we created individual data from publications which provided patient numbers by outcome score. Data were missing for a variety of trial types (acute/rehabilitation/stroke unit) and sizes, and functional outcome measure (mRS/BI), so it is unlikely that a systematic bias was introduced into the findings; however, the precision of the results may have been attenuated by the missing trials.

In summary, we suggest that trialists designing future stroke studies of treatments which are likely to act uniformly across populations should consider analysing functional outcome using an ordinal method that retains the natural ordering of the outcome data; in doing so, they will be able to maintain study power for a smaller sample size which will reduce the complexity (less centres), length and cost of trials. However, trials of thrombolysis (or other interventions where a likely asymmetrical hazard will be present alongside a symmetrical efficacy) should probably use current approaches which combine outcomes; in this respect, the decision to use excellent (mRS 0,1/2-6 [9]), good (mRS 0-2/3-6 [20]) or moderate (mRS 0-3/4-6 [21]) splits in functional outcome will depend on the expected severity of patients. Nevertheless, it is apparent that there is no perfect method for calculating sample size for stroke trials and other factors related to trial design and patient type should be considered. Software is available to calculate sample size using the approaches tested here.[11, 22]

REFERENCES

1. Bath FJ, Owen VE, Bath PMW. Quality of full and final publications reporting acute stroke trials. A systematic review. *Stroke*. 1998;29:2203-2210
2. The Optimising Analysis of Stroke Trials (OAST) Collaboration. Can we improve the statistical analysis of stroke trials? Statistical re-analysis of functional outcomes in stroke trials. *Stroke*. 2007;38:1911-1915
3. Rankin J. Cerebral vascular accidents in patients over the age of 60. 2. Prognosis. *Scottish Medical Journal*. 1957;2:200-215
4. The CONSORT Statement. Improving the quality of reporting of randomized controlled trials. *JAMA*. 1996;276:637-639
5. Gardner MJ, Altman DG. Statistics with confidence. 1989
6. Weaver CS, Leonardi-Bee J, Bath-Hexall FJ, Bath PMW. Sample size calculations in acute stroke trials: A systematic review of their reporting, characteristics, and relationship with outcome. *Stroke*. 2004;35:1216-1224
7. Mahoney FI, Barthel DW. Functional evaluation: The barthel index. *Maryland State Medical Journal*. 1965:61-65
8. Lindley RI, Waddell F, Livingstone M, Sandercock P, Dennis MS, Slattery J, Smith B, Warlow C. Can simple questions assess outcome after stroke? *Cerebrovascular Diseases*. 1994;4:314-324
9. The National Institute of Neurological Disorders and Stroke rt-PA Stroke Study Group. Tissue plasminogen activator for acute stroke. *New England Journal of Medicine*. 1995;333:1581-1587
10. Duncan DB. Multiple range and multiple f tests. *Biometrics* 1955;11:1-42
11. Whitehead J. Sample-size calculations for ordered categorical-data *Statistics in Medicine*. 1993;12:2257-2271
12. Payne R. Ssigntest *Genstat Reference Manual*. 1993;3
13. The Edaravone Acute Brain Infarction Study Group (Chair: Eiichi Otomo MD). Effect of a novel free radical scavenger, edaravone (mci-186), on acute brain infarction. *Cerebrovascular Diseases*. 2003;15:222-229
14. International Stroke Trial Collaborative Group. The international stroke trial (ist); a randomised trial of aspirin, subcutaneous heparin, both, or neither among 19435 patients with acute ischaemic stroke. *Lancet*. 1997;349:1569-1581
15. Furlan A, Higashida R, Wechsler L, Gent M, Rowley H, Kase C, Pessin M, Ahuja A, Callahan F, Clark WM, Silver F, Rivera F. Intra-arterial prourokinase for acute ischemic stroke. The proact ii study: A randomized trial. *Journal of the American Medical Association*. 1999;282:2003-2011
16. Song F, Jerosch-Herold C, Holland R, Drachler Mde L, Harvey I. Statistical methods for analysing barthel scores in trials of post stroke interventions: A review and computer simulations. *Clinical Rehabilitation*. 2006;20:347-356
17. Walters SJ, Campbell MJ, Lall R. Design and analysis of trials with quality of life as on outcome: A practical guide. *Journal of Biopharmaceutical statistics*. 2001;11:155-176
18. Wardlaw JM, Zoppo G, Yamaguchi T, Berge E. Thrombolysis for acute ischaemic stroke. *Cochrane Database Systematic Review*. 2003:CD000213
19. Mayer SA, Brun NC, Broderick J, Diringer MN, Davis S, Skolnick BE, Steiner T. The fast trial: Main results. *European Stroke Conference*. 2007
20. Hacke W, Markku K, Fieschi C, von Kummer R, Davalos A, Meier D, Larrue V, Bluhmki E, Davis S, Donnan G, Schneider D, Diez-Tejedor E, Trouillas P. Randomised double-blind placebo-controlled trial of thrombolytic therapy with intravenous alteplase in acute ischaemic stroke (ecass ii). *Lancet*. 1998;352:1245-1251

21. The FOOD Trial Collaboration. Effect of timing and method of enteral tube feeding for dysphagic stroke patients (food): A multicentre randomised controlled trial. *Lancet*. 2005;365:764-772
22. GenStat. Genstat release 8.1 (pc/windows xp),. 2005
23. Bland M. *An introduction to medical statistics*. Oxford: Oxford University Press; 2000.
24. Fligner MA, Policello GE. Robust rank procedures for the behrens-fisher problem. *Journal of the American Statistics Association*. 1981;76:162-168

TABLE 1

Comparison of sample sizes produced by 5 methods. Lower ranks imply the method produces lower sample sizes. Analysis by two-way ANOVA and Duncan's multiple range test; tests joined by the same band are not significantly different from each other at $p<0.05$.

Method	Mean rank	n	Banding
Comparing ordinal data (Whitehead) [11]	2.15	53	
Comparing means	2.28	55	
Comparing proportions (good outcome)	3.18	55	
Comparing proportions (excellent outcome)	3.37	54	
Comparing medians (Payne) [12]	3.92	54	

TABLE 2

Comparison of sample sizes using 4 methods of calculation relative to the proportion method for a good outcome (modified Rankin Scale $\leq$2 or Barthel Index $\geq$60) with results subcategorised by type of intervention. Data are median (inter-quartile range) multiplier.

Intervention	Trials n	Ordinal	Means	Proportion (excellent)	Medians
Thrombolysis	10	1.22 (0.73, 2.15)	1.36 (0.52, 38.84)	1.92 (0.43, 6.70)	2.06 (1.22, 3.35)
Anticoagulation	3	0.97 (0.59, 1.08)	1.03 (0.55, 1.03)	0.57 (0.16, 1.47)	1.64 (1.01, 1.78)
Antihypertensive	4	0.42 (0.34, 1.29)	0.88 (0.35, 2.43)	0.83 (0.01, 8.72)	0.27 (0.53, 1.98)
Antiplatelet	4	0.51 (0.28, 0.67)	0.48 (0.38, 0.66)	0.83 (0.35, 1.03)	0.82 (0.44, 1.11)
Feeding	1	0.07 (- , -)	0.04 (- , -)	0.11 (- , -)	0.14 (- , -)
Neuroprotection	17	0.71 (0.22, 1.09)	0.70 (0.42, 0.92)	0.92 (0.23, 2.34)	1.08 (0.41, 1.43)
Occupational therapy	7	0.44 (0.04, 2.20)	0.37 (0.03, 3.46)	0.30 (0.07, 20.77)	0.73 (0.06, 3.38)
Procoagulant	1	0.79 (- , -)	0.68 (- , -)	1.06 (- , -)	1.17 (- , -)
Stroke unit	8	0.88 (0.35, 24.32)	0.96 (0.22, 4.21)	4.36 (1.75, 31.82)	1.36 (0.56, 5.53)
Total	55	0.72 (0.47, 0.86)	0.70 (0.55, 0.94)	0.99 (0.71, 1.79)	1.12 (0.80, 1.40)

FIGURE LEGENDS

FIGURE 1a

Sample size comparisons at varying levels (β) of power for the IST trial of aspirin.

FIGURE 1b

Sample size comparisons at varying levels (β) of power for a trial of edaravone.

FIGURE 1c

Sample size comparisons at varying levels (β) of power for the PROACT II trial of intra-arterial prourokinase.

FIGURE 1a

Sample size comparisons at varying levels (β) of power for the IST trial of aspirin.[14]

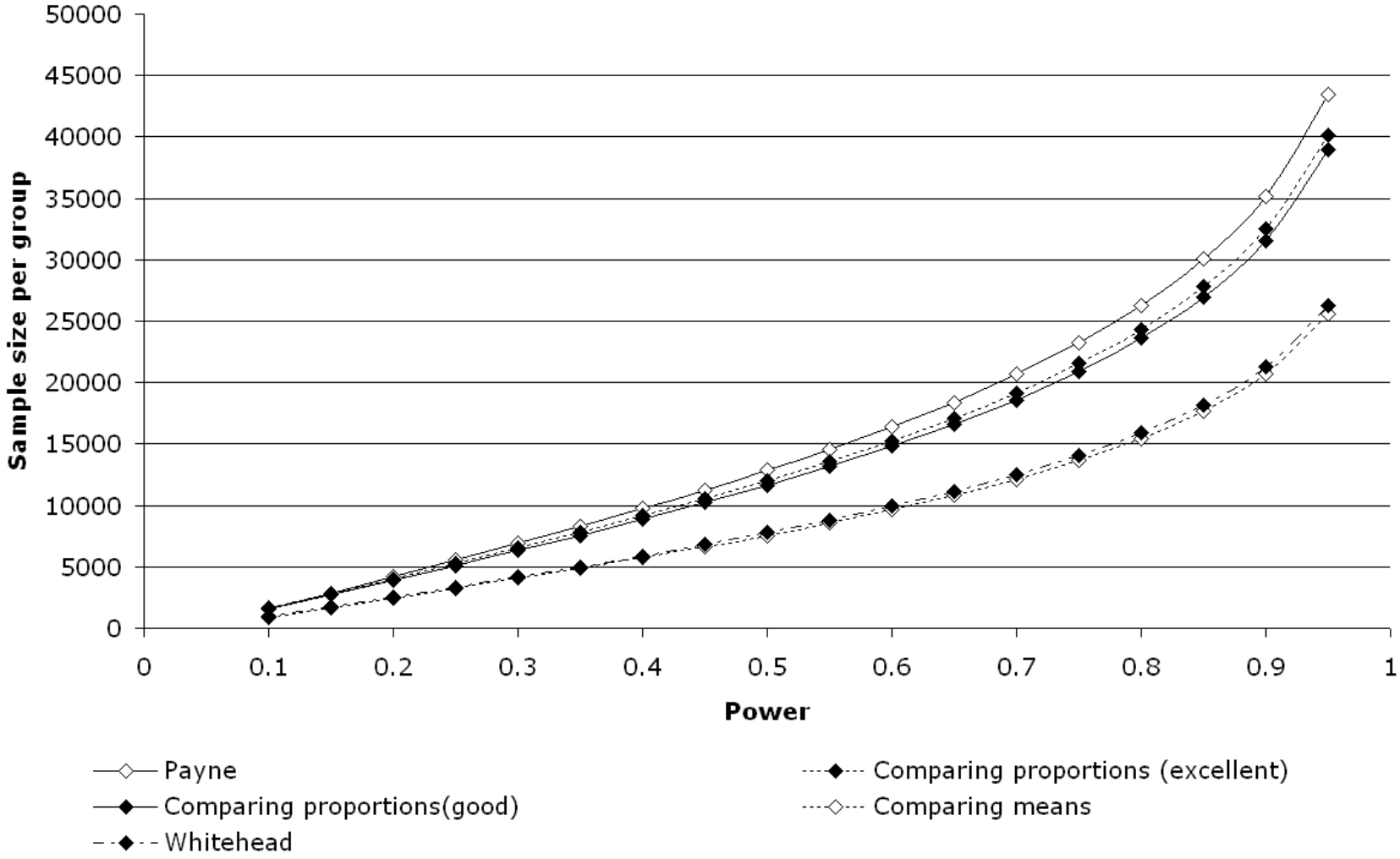


FIGURE 1b

Sample size comparisons at varying levels (β) of power for a trial of edaravone.[13]

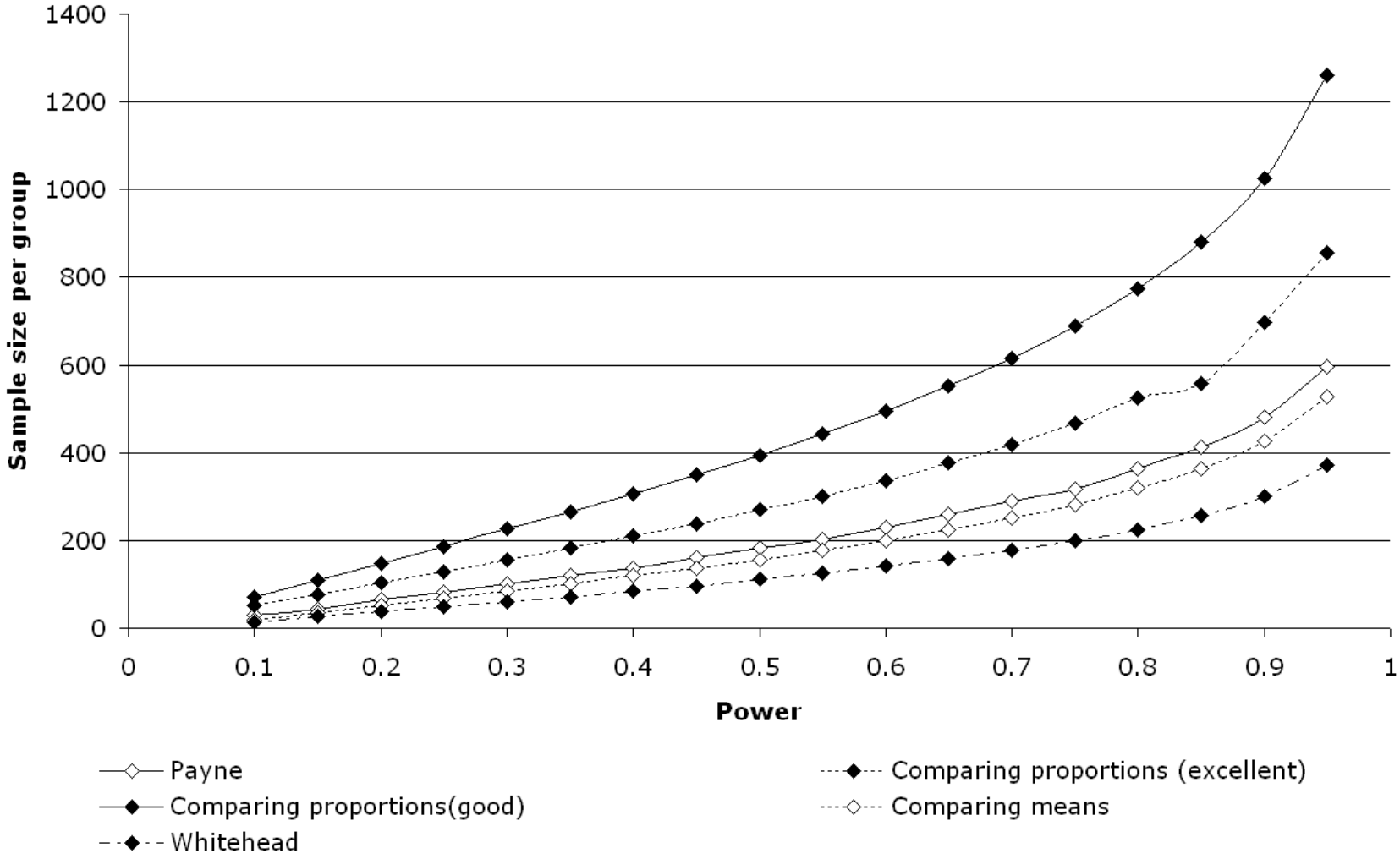
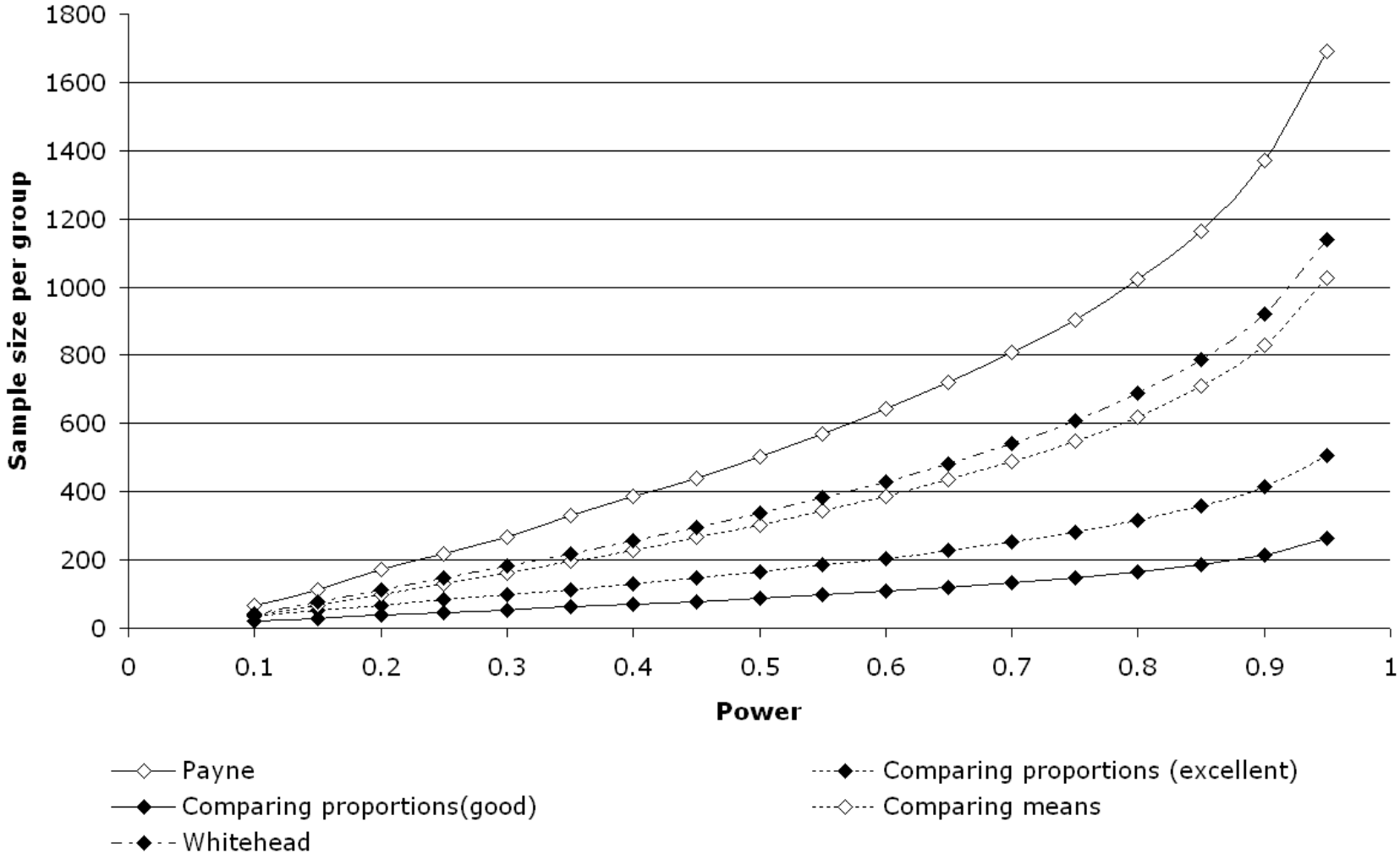


FIGURE 1c

Sample size comparisons at varying levels (β) of power for the PROACT II trial of intra-arterial prourokinase.[15]



APPENDIX 1

Sample size calculation

Comparing two proportions

The formula for estimating the sample size when the outcome is binary is:

$$n = \frac{(z_\alpha + z_\beta)^2 (p_1(1-p_1) + p_2(1-p_2))}{(p_1 - p_2)^2} \quad (1)$$

where n is the number of patients required in each group, p_1 and p_2 are the proportions of interest in the two treatment groups.[6]

Comparing two means

If a trial has an outcome which is continuous then the investigator may choose a comparison of means as the method of analysis for the primary outcome, e.g. using the student's t test. The appropriate sample size calculation is based on:

$$n = \frac{2\sigma^2 (z_\alpha + z_\beta)^2}{(\mu_2 - \mu_1)^2} \quad (2)$$

where μ_1 and μ_2 are the expected means in the two treatment groups and σ is the overall expected standard deviation.[23]

Comparing two medians

This method of sample size estimation for comparing ordinal data was proposed by Payne [12] as part of the Genstat [22] statistical program and is relevant when the Wilcoxon test or the robust rank test [24] will be used to analyse the primary outcome once the trial is completed. The method calculates an approximate sample size needed based on the probability of response (i.e. the probability that an observation in one sample will be greater than the equivalent observation in the other sample) that should be detectable by initially assuming a Normal approximation.

This is then refined by calculating powers for a range of replications centred around that approximation.[12]

Comparing ordinal data

Sample size estimation for comparing two groups of ordinal data using the technique of ordinal regression was proposed by Whitehead.[11] An estimate of the expected odds ratio and proportion of patients expected to fall into each category on the scale being used for one of the treatment groups is required. The sample size per group is given by:

$$n = \frac{6\left[(z_\alpha + z_\beta)^2 / (LogOR)^2\right]}{\left[1 - \sum_{i=1}^{k} \bar{\pi}^3\right]} \qquad (5)$$

where OR is the odds ratio of being in category i or less for one treatment group compared to another, k is the number of categories on the scale of interest, and $\bar{\pi}$ is the mean proportion of patients expected in category i.

All sample size formulas used are asymptotic large-sample formulas that assume convergence to a standard normal distribution.